Transcription process, L2 essays

Google Doc version
(https://docs.google.com/document/d/1bAXIK85ttWJuOXRmsSm_-fTdAlMNKpJG7ddgIrHB77o/edit#)

Elena Volodina, Bea Megyesi, June, 04 - July, 15, 2018;

Draft

Transcription flow

This document contains instructions for manual conversion of hand-written essays to a digital format, and recommends the following flow:

1. Acquaintance with guidelines (annotators)
2. Transcription workshop (annotators + researchers)
3. Transcription (individual annotators)
4. Cross-consultation, in uncertain cases (annotator-annotator or annotator-researcher)
5. Transcription check (third party)

Acquaintance with the guidelines (1) means actual study of this document, optimally combined with a practical test-case using a number of real-life essays, to see how different questions can be guided.

Transcription workshop (2) is a practical one-day session when several annotators work on actual essays and discuss uncertain cases between themselves and with a responsible researcher. The workshop is aimed at resolving subjective judgements in favour of objective decisions. Optimally, annotators involved in this process can build some network so that when uncertainties arise during their later work, they can ask each other.

Transcription, individual phase (3) is an individual process when each annotator is working on his/her share of essays.

Cross-consultation (4) is a step which can be of use in uncertain cases. We recommend to get in contact with another assistant or a responsible researcher to double-check any uncertainties.

Transcription check (5) is performed by a third party, e.g. another annotator. During this stage random checks are performed on the transcribed files.

Transcription guidelines

Two **major principles** for essay transcription are: * not to correct author's mistakes and * not to make personal assumptions.

The following **rules** should apply to transcription of hand-written essays:

Authenticity of writing

Errors should be preserved from the hand-writing, e.g. *no error correction!* (see in Figure 1). Tips: disable spell-checker.

If there is a dubious case - for instance, whether the learner has written correctly or incorrectly - a *positive assumption* should be made. For example, if it is unclear whether two words have been written as one or as two items (with too little space between them), the “positive” assumption would be that the learner meant to write two items, and in practice in the transcribed format the string should be separated into two words. When it is obvious that the two words are written as one, that should be preserved (typed as one item). (See in Figure 1)

Handwriting

In many cases some basic knowledge of Swedish should help to understand what is written (see in Figure 1; as well as Figure 2, lilac underlinings). If the handwriting is illegible, write $ (dollar-sign) for each character that cannot be understood.

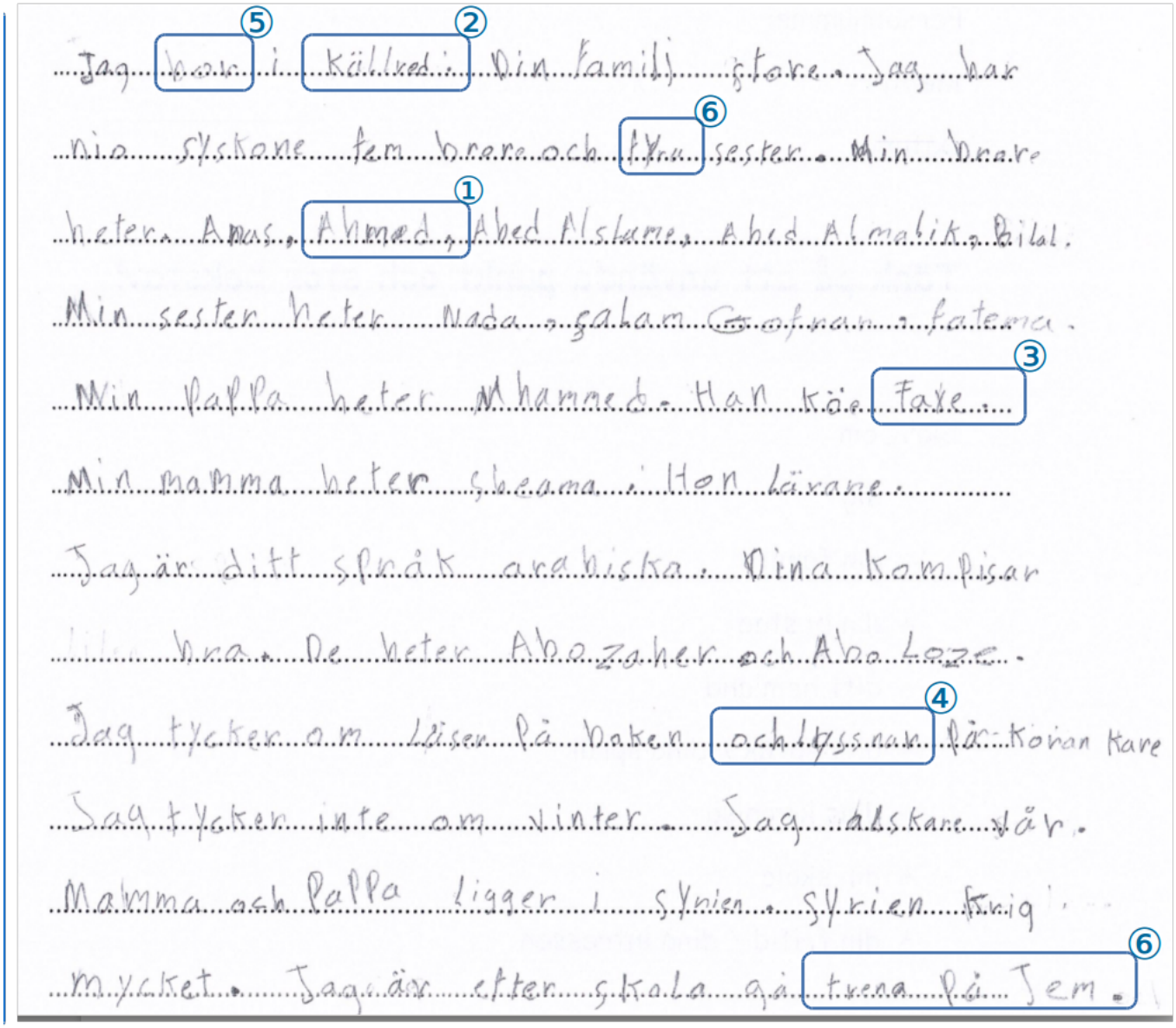

Figure 1. Essay example. Level A1 (beginner), nr tokens: 117, topic: Presentation/Om mig; transcription time: 15 min.

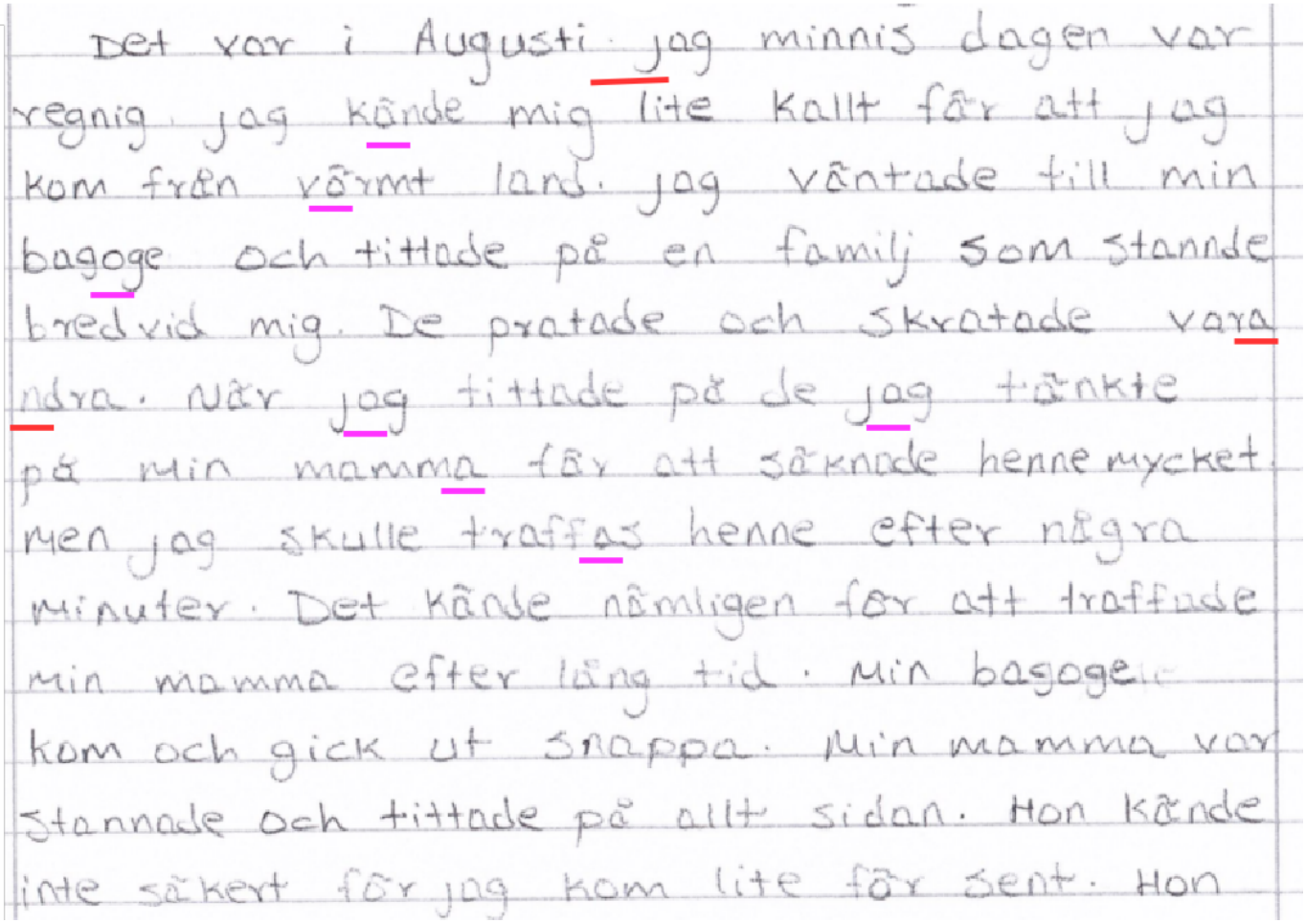

Figure 2. Deciphering letters in student writing: o vs a; capital/non-capital; hyphenation. Level B1 (intermed.), nr tokens: 330, topic: Min första dag i Sverige; transcription time: 34 min.

Non-existent letters

Sometimes students can be very creative and “invent” their own letters (see Figure 3). In the case when you know there is an equivalent letter in another language, use that one, as for example, in the first word in Fig.3: *Sverigë*.

In the case, if there is no way to reflect that letter in writing, choose the closest one in shape, keeping to the positive assumption described in 5. For example, in the second example in Fig.3, the options could be o and å, but å holds the positive assumption, så the transcribed version should contain the word *går*.

If there is no way you can report a corresponding "created" letter, use dollar-sign $ as if it is an intelligeble letter.

Figure 3. Invented letters

Graphical issues (supra-linguistic features)

1. Insert line breaks (“Enter”) to introduce new paragraphs.
2. Differentiate between capital letters and non-capital ones (see □ in Figure 1).
3. Keep smileys
4. If a student has stricken out some text, that text should not be digitized.
5. Don’t bother about leaving indentation. We are primarily interested in the contents and the language, not in the graphical reproducibility.
6. All edits, like underlinings, are discarded. We keep only the text.
7. Comments in the margin are a case for interpretation. If it is obvious that the margin text is a part of the running text, it should be added into the essay. Otherwise - not. (We need, though, to make transcribed versions comparable to the digitally-born essays. Here, we need to see whether there are any cases where students have left their comments in the digitally-born versions, e.g. in the format of footnotes.)

Rule of thumb

Re-read each hand-written essay once again and compare with your transcribed version. You will get used to the student’s handwriting by then, and will - probably - understand better what is written. Another reason for re-reading the essays is to double-check that no unintended error-correction is introduced (rules 1-5).

Time estimation

To be able to give some time estimation, we have taken time for digitization of several essays per level. The summary follows in table 1 below.

It will take longer per essay in the beginning, when annotators are not yet confident with the guidelines and the process itself. The time will also depend upon the legibility of handwriting, and challenges of the writing, i.e. presence of challenging interpretations/uncertainties. Take the time estimations below only as an approximation.

Average	A1	A2	B1	B2	C1
Characters / essay	459	702	1761	-	-
Words / essay	85	134	331		
Minutes / essay	10	9	24,5		
Words / minute	8,25	15	13,5		

Table 1. Time estimation essay transcription at different levels

During your work, write down your time per essay in an excel sheet acc. to the example below (Table 2):

Essay-ID	Level	L1	Nr words	Time in minutes	Comment, if necessary

Table 2. Time estimation for annotator work on transcription

Text format / use of a kiosk version of SVALA

The work on hand-written essays is potentially risky, since certain amount of personal information in the text (as well as handwriting itself) may give away a person behind the text. That is why this work has to be performed in a safe environment. You will be introduced to the SweLL “kiosk” option for that. See instructions for "kiosk" here: https://github.com/spraakbanken/swell-project/blob/master/SweLL_kiosk_user_manual.md

You can save your work in any format while you are working, but your should deliver it in a plain text format (in unicode utf-8).

Save information about time used for digitizing an essay - for our statistics.

List of questions

1. Essay metadata visible in *kiosk* computer while transcribing?
2. Pdf-viewer or Picture-viewer in *kiosk* computer?
3. Text editor & Virtual keybord for various tokens in *kiosk* computer?
4. At the moment we are focusing only on the case of hand-written essays. Should we cover even

OCR of digital materials, e.g. pdf of a digitally-born essay? In OCR-ed pdfs the text should be extracted and manually checked/corrected for OCR errors. (Comment: Vet inte om vi får några sådana. Behöver först testa hur bra OCR-kvaliten blir, samt vilka program man ska använda för att extrahera text från pdf-filer & hur lång tid det tar att gå igenom och korrigera dessa. På Nationella prov OCR-ar man detta.)